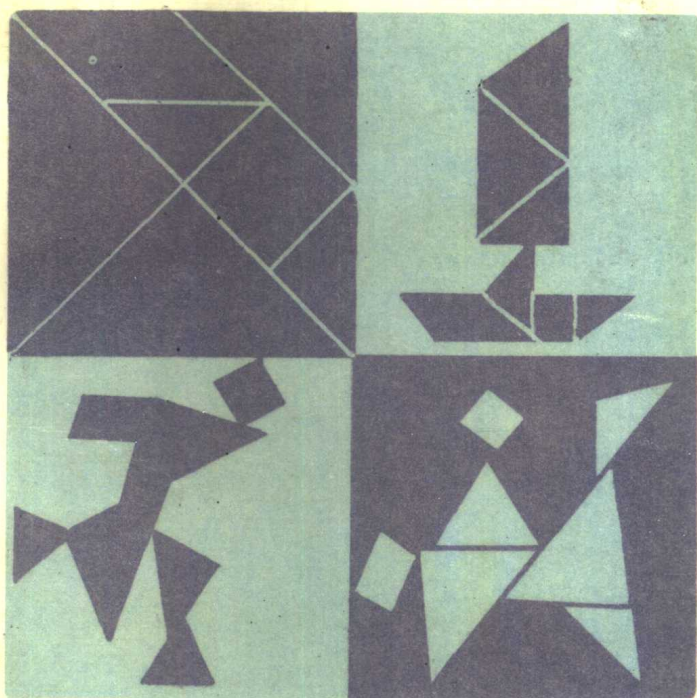


高等学校文科教材

教育测量学

王汉澜 主编



河南大学出版社

高等学校文科教材

教育测量学

王汉澜 主编

河南大学出版社

教育测量学

王汉澜 主编

责任编辑 史锡平

河南大学出版社出版

河南新郑印刷厂印刷

河南省新华书店发行

850 × 1168毫米 32开本 14.625印张 367千字

1987年9月第一版 1987年9月第一次印刷

1—3,000册

ISBN 7-81018-031-2/G·16

统一书号: 7435·041 定价: 2.80元

前 言

我国的教育目的是使受教育者在德、智、体等方面全面发展，成为有理想、有道德、有文化、有纪律的社会主义建设人才。我们学生学业成绩的考核，学校教育质量的评定，以及社会各部门的招工用人，都应实行德智体全面考核。而如何使这些考核工作科学化，就需要有一套科学的方法。特别是随着教育体制改革的深入发展，教育管理要科学化，教育评估将普遍展开，对教育的研究要作定性和定量的分析，对现行的考核制度要进一步使其完善，所有这些都需要借助于教育测量学的知识。

教育测量学就是研究如何测量学生的学业成绩、学习能力、兴趣爱好、思想品德以及教育上许多措施的效益等问题的一门科学，它是改进教学、实行科学化教育管理和教育研究的重要工具。而在目前，我国关于教育测量学的书籍还很少，有些高等学校虽已开设教育测量学，但还缺乏这方面的教材，我们受国家教育委员会的委托编写了这本教育测量学。

本教材分四编十三章，第一编是论述测量的基本原理，第二编是介绍测量的一般步骤和方法，第三编是分述各类教育测验，第四编是研究测量结果的统计处理，其目的是使学生比较系统而又全面地掌握教育测量的基本理论和方法。在编写中，我们力求理论联系实际，作到简明、准确、系统、完整，适合教学之用。

本书由河南大学教育系教授王汉澜担任主编，并执笔第一章，第二、三、九章由西南师范大学高等教育研究所教授杨宗义执笔，第四、五、六、七章由山东师范大学教育系副教授赵承福

执笔，第八章由河南大学教育系讲师丁秀峰执笔，第十、十一章由南京师范大学教育系副教授班华执笔，第十二、十三章由河南大学教育系讲师扈涛执笔。在编写过程中，我们充分发扬了民主，集思广益，进行了认真的讨论。但是，由于我们的水平所限，编写时间仓促，错误之处在所难免，敬希读者批评指正！

编 者

1986年12月

目 录

第一编 测量的基本原理

第一章 教育测量概述

- 第一节 教育测量的含义与特点…………… (1)
- 第二节 教育测量发展的历史…………… (6)
- 第三节 教育测量的要素和种类…………… (15)
- 第四节 教育测量的功能和对教育测量应有的认识
与态度…………… (21)

第二章 测验的信度

- 第一节 信度的意义…………… (26)
- 第二节 信度系数的计算…………… (35)
- 第三节 提高测验信度的方法…………… (58)

第三章 测验的效度

- 第一节 效度的意义…………… (68)
- 第二节 效度的估计…………… (73)
- 第三节 提高测验效度的方法…………… (90)

第四章 测验的难度与区分度

- 第一节 难度及其计算…………… (96)
- 第二节 区分度及其计算…………… (108)

第二编 测量的一般步骤和方法

第五章 测验的编制与实施

- 第一节 测验目的的确定…………… (119)
- 第二节 试题的编制…………… (125)

第三节	试题的评分	(141)
第四节	试卷编辑与测验实施	(147)
第六章 测验项目的质量分析		
第一节	测验质量分析的意义	(154)
第二节	常模参照性测验的项目分析	(158)
第三节	目标参照性测验的项目分析	(163)
第四节	多重选择题的项目分析与改进	(167)
第七章 标准化测验的基本程序与应用		
第一节	标准化测验的特点	(173)
第二节	标准化测验的基本程序	(175)
第三节	标准化测验的应用	(189)

第三编 各类教育测验

第八章 学科测验

第一节	学科测验的意义和编制学科测验应注意的问题	(193)
第二节	语文测验	(200)
第三节	数学测验	(223)
第四节	学科测验结果的解释	(233)

第九章 智力测验

第一节	智力测验的意义	(237)
第二节	智力测验的种类	(247)
第三节	几种重要的智力量表	(256)
第四节	智力测验结果的解释	(270)

第十章 人格测量

第一节	人格测量的意义	(281)
第二节	自陈量表法	(290)
第三节	投射法	(309)

第十一章 品德测量

- 第一节 品德测量的意义.....(317)
- 第二节 情境测量法.....(320)
- 第三节 行为观察法.....(324)
- 第四节 自编问卷测量法.....(328)
- 第五节 谈话法.....(335)
- 第六节 评定量表法.....(337)
- 第七节 社会关系的测量法.....(356)

第四编 测量结果的统计处理

第十二章 教育测量分数的整理和转换

- 第一节 教育测量分数的整理.....(363)
- 第二节 教育测量分数的转换.....(374)
- 第三节 多科测验分数加权评定标准化.....(384)

第十三章 测验分数的检验

- 第一节 区间估计.....(394)
- 第二节 Z检验.....(400)
- 第三节 t检验.....(405)
- 第四节 χ^2 检验.....(414)

附：利用电子计算机进行测验分数的整理、
转换和计算程序

附表

- 附表1 相关系数(r)显著性临界值表
- 附表2 正态曲线下Z为正值概率
- 附表3 χ^2 的数值表
- 附表4 t值表
- 附表5 F值表

本书主要参考文献

第一章 教育测量概述

第一节 教育测量的含义与特点

一、教育测量的含义

测量 (Measurement) , 通常是指人们对客观事物进行某种数量化的测定。如人们用天平、磅、秤测定物体的重量, 用尺子测定物体的长宽, 用温度计测试人体的温度等等, 人们凭借着测量, 才能对客观事物进行数量化的确定, 从而更好地认识事物和对事物进行比较和分析。测量通常是以数量来表述其结果的, 没有数量来表示所得的结果, 如说“这个物体是重的”、“那个物体是轻的”、“这个人高”、“那个人低”, 这都不能称为测量。

随着科学技术的发展, 人们不但对物体的长度、重量、温度以及时间、空间、运动等物理特性做出了越来越精确的测量, 而且对人的知识、技能、智力、兴趣、气质、性格、品德等精神特性也能进行测量。从而加深了对人类精神现象的了解, 促进了教育科学的发展。

教育测量 (Educational Measurement) 就是对学生们的学习能力、学业成绩、兴趣爱好、思想品德以及教育措施上许多问题的数量化测定。它主要用于对学生精神特性的测定。

对学生精神特性的测量比对物体的物理特性的测量要困难得多, 因为学生是活生生的人, 学生的精神时刻在发生变化, 不易控制。一块钢铁, 今天重1200克, 过了几天或几个月, 仍旧是1200克。一幅布, 今天长1.5米, 过了几个月, 仍旧长1.5米。但是, 一个学生今天能在一小时内做9道数学题, 到了明天, 在一小时内, 也许能做8道或10道数学题。因为一个人的能力或作业成绩

极易受作业题的难度和个人的休息、睡眠、注意力的集中程度、心情等等的影响，是时刻变动、不易控制的。而且测量物理性质的单位(克、米、温度、时间等)比较精确、稳定，可以采用直接的方法，称出物体是多重，量出物体是多长。对学生的精神特性的测量，就不能用一把尺子或一架天平直接地把他的智力和演算能力测量出来，只能通过他的动作，间接地测量出他的智力和演算能力。测量的结果，也不象物理特性的测量那样精确、稳定。

但是，这并不等于说，学生的学习能力、学业成绩和思想品德等是不能测量的。美国心理学家桑戴克(E·L·Thorndike 1874—1949)在1918年曾说：“凡物之存在必有其数量”，美国测量学家麦柯尔(W·A·Mccall)于1922年又提出：“凡有数量的东西都可以测量。”^①这即是说，世界上的事物现象都有程度上的不同，程度上的不同也就是数量上的差异，既然有数量上的差异，就可以进行数量化的测定。学生的学业成绩有好有坏，有的学生能回答较多的问题，有的同学能回答的问题比较少；学生的思想品德、兴趣爱好等也有不同，有的学生在某种情况下，他的某种思想行为比别的学生表现的较多、较明显、较强烈、较快一些，或者较少、较差、较慢一些。这些程度上的差异，就为教育测量提供了可能和理论基础。同时，学生的心情也是可以运用仪器，从脉搏、呼吸、血压、心跳、皮肤电和外部行为的表现中测量出其差异程度的。虽然，在教育测量上不如物理特性的测量那样精确、易行，但绝非不可测量的。

质言之，无论是在天文领域中的测量，或是在物理、生理、心理、教育等领域中的测量，都是没有绝对精确和毫无误差的。测量者虽力求作到精确无误，但也只能使测量达到一定限度的精确，不能完全消除误差。这是因为有三种情况，都会导致测量的误差：

1、测量的仪器和设备。测量工具的精密程度会影响测量的

① 冻选善：《教育测验》，商务印书馆，第2—3页。

精确程度，测量时使用的仪器和设备条件越好，误差就会越小，而绝对精密、完美无缺的仪器和设备，是难以制造和得到的。

2、所测量的事物缺乏一致性。所测量的事物是否具有一致性，这是影响测量结果是否精确的一个重要原因。如测一根树藤的拉力，由于选取的树藤的部位不同，所测得的拉力大小会不相等；在一堆矿石中，由于选取的矿石块不同，所测得的含有某种元素的成分也会有异。

3、测量人的因素。由于测量人的技术不熟练，或观察能力不强，一时注意力不够集中等，都会导致测量的不精确。

在教育测量中，同样存在着上述三种情况的误差。教育测量中所使用的工具绝不会精密无缺，测量的量表很难学生的知识、才能、思想一点不漏的全部测量出来，学生的身心在不同的时候、不同的情况下有不同的表现，教育测量所能控制的条件是有限的，因而，教育测量是有误差存在的，教育测量的精确程度是不如物理特性的测量的。正因为这样，我们在教育测量工作中，应该采取有效的手段，科学的编制量表，妥善的选取和控制测量的对象，科学的进行评分，尽力减少测量的误差，提高测量的精确程度，并且还要周密而严谨地表述测量的结果。譬如，一个具有测量学和统计学知识的人，在介绍某一学生的测验成绩分数时，同时把该次测验的团体分数平均数、标准差和差异性检验情况，也说出来，这样，就使测量结果的表述，更为科学。随着生理学、心理学、统计学、电子学的发展，教育测量的技术和方法在不断地完善，教育测量学必将成为一门更加精确的科学。

在教育工作中，经常使用测验、考试、评价三个术语，这三个术语与测量既有联系，又有区别。

测验 (Test) 是引起某种行为的工具，通常是指运用某些仪器、试题来引起人们的某种行为，从而测定人们的某种特性。它是进行数量化分析和科学推断的前提和手段。测量比测验的含义

要广泛些，测量不仅包括着运用仪器、试题来进行测定事物的质量与特性，而且还包括着运用调查、观察等方法来测量事物的质量与特性，如通过图书借阅情况的调查或对学生活动的观察，来测量学生的兴趣与爱好。测量既包括对事物的测验，又包含对事物进行数量化的分析，并对测验结果进行一定的解释和评价。不过，在已往测量书籍中，测验与测量常互相通用。

测验与考试 (Examination) 也不尽相同。我们平时所说的考试，通常只凭教师自己的经验去出试题和评分，带有主观随意性。测验是经过较细致的科学分析才编制出测验的题目，在测验的程序和评分方面也有较严格的要求。考试一般用于考核学生的学业成绩，测验不仅用于考核学生的学业成绩，还使用于心理特性的测量。

测量与评价 (Evaluation) 既有联系，也有区别。测量的数量化结果，如果不依据测量的目的进行分析、解释和评价，就是无意义的东西，所以测量包含有一定的评价。不过，测量强调的是数量化，一般要尽可能地减少判断的作用。评价则是突出价值观，充分重视对问题的分析与评断。它要根据测量的结果，评断是否达到了教育的目的和要求，达到了什么程度，甚至评价教育的措施是否符合社会的需要？测量的结果是评价的重要依据，而不是评价的唯一依据，评价需要依据更多的方法和更多的资料来进行。它含有综合观察、实验、调查、测量、阅读文献等方面获得的资料，对某一问题进行评估的意思。如对学生毕业成绩的评价，除依据测量所得的分数外，还要依据学生的平时作业、课堂回答、过往的学习成绩等情况。

教育测量与心理测量有着密切的内在联系。心理测量是测量人的心理活动的差异，一般包括智力测验、人格测验等。教育上运用的心理测量，是测量学生在教育影响下的心理活动方面的差异。教育测量除包括智力测验、人格测验外，还包括学生成绩测

验。教育测量是着重于对教育效果的测量，测量学生在教育的影响下，学生心理上的变化情况，并常根据教育目的和教材内容的要求来编制量表，而心理测量的量表则不宜直接取材于教科书。

二、教育测量的特点

教育测量与物理特性的测量相比，具有以下一些特点：

（一）教育测量一般是间接测量

一个物体的重量、长度、温度，可以用天平、尺子、温度计直接测量。人的知识、智力、才能，是人的大脑活动的表现，科学发展到今天，我们还无法直接测量人的这些大脑活动情况，即不能直接测量人的心理实体。我们只能依据人的外显行为，间接的测量人的心理活动的特点与水平。我们只能通过学生对测验题目的反应和一些行为表现，运用推理、判断的方法，来间接的测量出他的知识水平、智力高低和品德好坏。在教育测量中，除了对学生身体的素质和发育的测量可以直接测量外，其余则是间接测量。

（二）教育测量的度量单位是相对的

物理特性测量的度量单位是绝对的，一米就是一米，一克就是一克，无论是在上海，还是在北京，都是相等的。100克比95克重5克，50克比45克重5克。这两个相差的5克是相等的。当然，物理特性测量的度量单位的绝对性，也是有条件的，也具有一定的相对性。如在海平面上—磅重的物体，在高山顶上就会不到—磅重。钢刻度尺在100℃时就比在10℃时长一些。不过，在一般的情况下，物理特性测量的度量单位是比较稳定的、绝对的。

在教育测量中，一个学生在一次测验中得70分，可能在这个班上较高的分数，而在别的班、别的学校可能是较差的分数；甲在某校的70分并不一定比另一学校乙的80分低，因为考试的题目难易不同，评分标准不同，班上学生成绩分数的分布情况不同，因而分数的价值并不相等。再如，在百分制的评分方法中，

60分比55分多5分，90分比85分多5分，这两个相差的5分，并不是由相等的标准来确定的。一个在测验中得80分的学生知识量并不等于在同一测验中得40分学生知识量的两倍。因此，教育测量的度量单位是相对的。一个学生的学业成绩好坏、智力高低和能力大小等，都是就其在所在团体的整个分数序列、行为序列中的地位来说的，其测量的分数单位，并非绝对的。为了使教育测量的分数具有可靠性、可比性，必须根据标准测量的原理和方法，编制测验量表，制定科学的评分标准和方法，并将测量所得的原始分数转换为可以直接比较的量表分数（百分等级分数、标准分数、T分数等）。这将在本书的第十二章论述。

（三）教育测量是为实现教育目的服务的

教育测量既与物理特性测量不同，也与社会上的职业测量有所不同，它是为改进教育工作，提高教育质量，更好地实现教育目的服务的。教育测量的各种测验量表的制定，都必须根据教育目的的要求，以各种教材和教育内容为依据。教育测量的目的是为了掌握学生在学业、智力和思想品质等方面的情况，了解教育和教学的效果，以便更合理地组织教育活动，安排教学内容，选择教学方法，贯彻因材施教，促进学生在德、智、体等方面全面发展。它不能脱离教育目的和教材的要求，随意地制定测验量表，任意地进行测量。对各种教育测量结果的评价，也都应依据教育目的的要求来进行。那种不顾教育目的的要求，为测量而测量的作法，是不符合教育测量的要求的。

第二节 教育测量发展的历史

教育测量发展的历史，大体上可以分为三个阶段：

一、主观经验性考试阶段（1864年以前）

教育测量，由来已久，早在我国的西周时期，就初步建立了学校教育制度，实行了教育考评。根据《学记》的记载：“古之教者，家有塾，党有庠，术有序，国有学。比年入学，中年考校。一年视离经辨志，三年视敬业乐群，五年视博习亲师，七年视论学取友，谓之小成；九年知类通达，强立而不返，谓之大成。”到了隋炀帝大业二年（公元606年），我国开始实行了科举制，通过分科考试来选取人才，采用了贴经、墨义、策问、诗赋等方法去测试考生。这在历史上是一个伟大的创举，对世界文化有很大的影响。法国大革命时期，资产阶级启蒙思想家伏尔泰（Voltaire 1694——1778）曾赞叹地说：“人类精神，肯定想象不出比这样的政府更好的政府。在这个政府里，重要的衙门彼此统属，任何事情都在那里决定，而其成员，都是先经过几场严格的考试的。”^①欧美各国十八世纪末、十九世纪初采用考试的方法选拔官吏，是从我国学去的。欧洲学校的笔试开始于1702年英国的剑桥大学，1845年美国麻省波士顿市教育委员会才开始采用笔试的方法考查该市所属学校的毕业生。国内外不少学者认为，教育测量是起源于我国的科举考试制度。

教育测量虽起源于我国，由来已久，但是，我国古代的考试，一般是试题的数量少，知识的复盖面窄，不足代表所测学科知识的内容；而且没有客观的评分标准，阅卷者可以根据个人的知识、经验、爱好等来进行评分，评分易受阅卷者主观因素的影响，很难保证评分的客观性。因而，它不是科学的教育测量，需要改进。

二、客观标准化测验阶段（1864——1940年）

自十九世纪末到二十世纪三十年代，欧美资产阶级学者为了适应社会要求有效地培养人才，在教育科学化运动的推动下，对学校中的考试，进行了积极的研究，排除了试题编制的片面性和

^①转引自《教育统计与测验文选》，1984年第1期第55页。

评分方法上的主观经验性，使教育测量走上了客观性测验 (Objectivity test)。而后又经研究，使客观性测验标准化，成为客观标准化测验。所谓客观标准化测验 (standardiation test)，简单地说，就是根据一定的目的，认真细致地编制测验量表，科学的规定出标准答案和给分标准，先对大量的被试者进行测验，再按年龄或年级分别求出各年龄或各年级在测验中所得成绩分数的平均数和成绩分数分布情况〔即“常模”(Norm)〕，而后，以此为标准来衡量以后的被试者所得成绩。这种测验的发展经过是这样的：

1864年，英国格林威治医院(Groenwith Hospita School)一位教师名叫费舍(George Fisher)，他深感评定学生成绩，没有客观标准可循，仅凭教师主观判断，会使评定结果因人而异，他便努力收集许多学生成绩样本，分别优劣，汇集一本名叫《量表集》(Scale book)的书，作为度量学生各科成绩的标准。该书中列有评定各科成绩的等第，评定学生成绩时，可将学生的成绩与量表上的样本等级相对照，以确定学生成绩的等第。这可以说是客观标准化测量的萌芽。如此评定学生成绩，比没有标准可循，当然是前进了一步。但该《量表集》仅是根据费舍个人的思想和经验编制而成的，缺乏科学的依据和精密的分析。

美国来斯(J·M·Rice)博士主张用划一的测验去考查、比较各校学生的成绩，他于1895——1905年期间，编制了算术、拼字、语言等测验。其《算术测验》是用于考查四至八年级学生算术的能力，每级各有八题；其《拼字测验》是令学生默写他所编制表中的生字；其《语言测验》是将选定的文章，向学生朗诵一遍，命学生根据原文大意作成一文，视文章的语句及篇章结构确定分数。来斯用他编制的测验测量过数万名学生，并证明了拼字教学，由于教师教学方法的原因，在某些学校每日教30分钟，教授8年，并不比另一些学校每日教15分钟，教授8年的

成绩好。从而引起了许多教育家对教育测量的重视。

法国的比纳 (A · Binet, 1857——1911) 1898年在《科学》杂志上发表了一篇题为《个人心理学中的测量》，文中提到许多测验，如画方形、比较线的长短、记忆数目、词句重组、回答含有道德判断的问题、了解抽象文章的意义、折纸等等。这些测验方法后来都采用于他的智力量表中。在这篇文章中，他还提出了心理测量的根本原理，是在于将个人的行为与他人比较以归类。这是近代测量理论的基本思想。1905年比纳与其助手西蒙 (T · Simon, 1873——1961) 在《心理学年报》上发表了《诊断异常儿童智力的新方法》，在这篇文章中介绍的就是第一个智力量表——《比纳·西蒙量表》。该量表有30个由易到难排列的项目，可以测量出高低不同的智力。1908年比纳对其智力量表进行了第一次修订，把测量项目增加到59个，并采用了智力年龄的方法计算成绩。从3岁组到13岁组，儿童考试的结果，都可用智力年龄表示之。1911年比纳对其智力量表又作了第二次修改，增添了许多新测验，废除了一些旧测验。该年比纳去世。比纳、西蒙在心理和教育测量的发展史上，作出了不可磨灭的贡献，比纳被称为智力测量的鼻祖。

比纳·西蒙智力量表最完善的订正本是由斯坦福大学心理学教授推孟 (L · M · Terman, 1877——1956) 1916年发表的量表。推孟费了5年时间，经过严密的订正，才修正发表了这个新量表。该量表载于推孟所著《智力测量》一书中。新量表中共有90个测验项目，并引用了智力商数。智力商数是在心智商数的基础上提出的，心智商数原为德国心理学者斯特恩 (W · Stern) 在其1904年出版的《测量智力的心理方法》一书中提出的。心智商数的求法，是将智力年龄被实足年龄除，以心智商数表示聪明程度。推孟把心智商数改称为智力商数，并且为了避免小数起见，将商数乘100。其公式为：
$$\text{智力商数} = \frac{\text{智力年龄}}{\text{实足年龄}} \times 100$$